Regularization

전 재 욱

Embedded System 연구실
성균관대학교

Outline

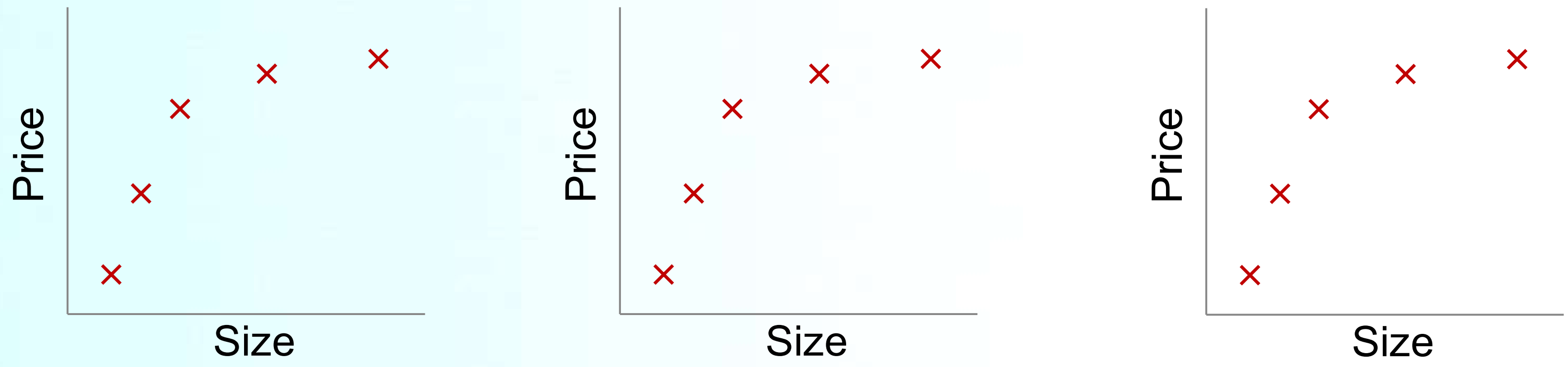
- Problem of overfitting
- Cost function
- Regularized linear regression
- Regularized logistic regression

Outline

- Problem of overfitting
- Cost function
- Regularized linear regression
- Regularized logistic regression

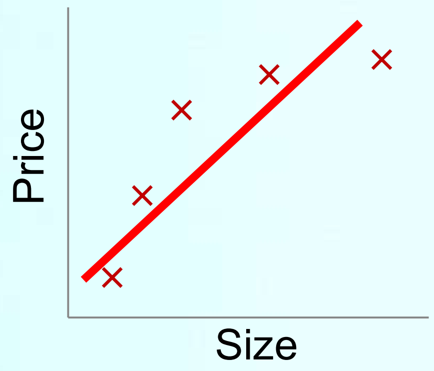
Overfitting

- Example
  - Linear regression (housing price)



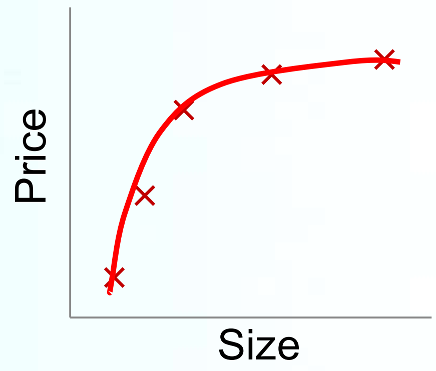
Overfitting

- Example
  - Linear regression (housing price)



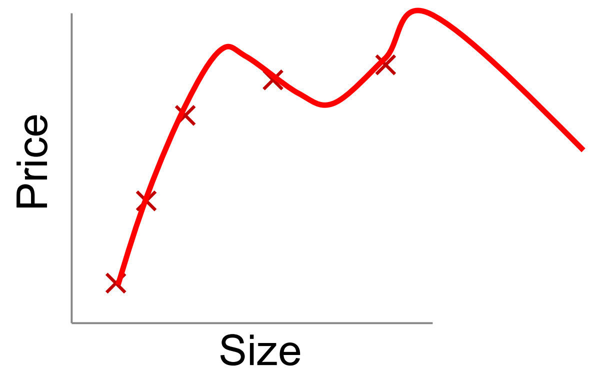
$\theta_0 + \theta_1 x$

“Underfit”
“High bias”



$\theta_0 + \theta_1 x + \theta_2 x^2$

“Just right”

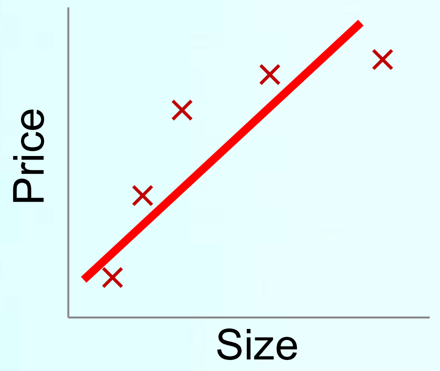


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

“Overfit”
“High variance”

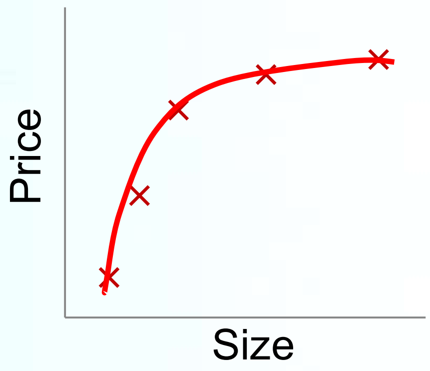
Overfitting

- Example
  - Linear regression (housing price)



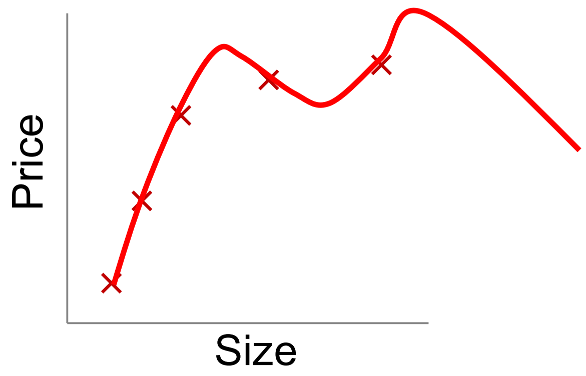
$$\theta_0 + \theta_1 x$$

"Underfit"
"High bias"



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

"Just right"



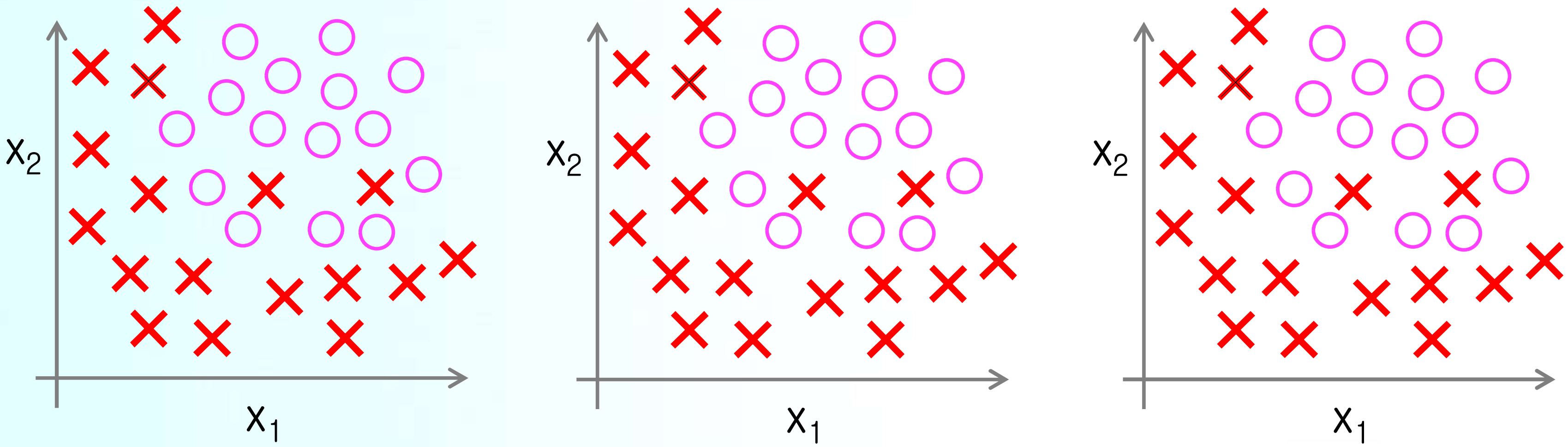
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

"Overfit"
"High variance"

- Overfitting
  - If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \approx 0$)
    - but fail to generalize to new examples (predict prices on new examples)

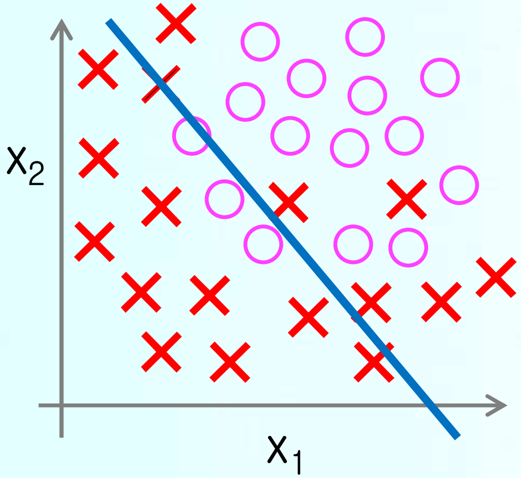
Overfitting

- Example
  - Logistic regression



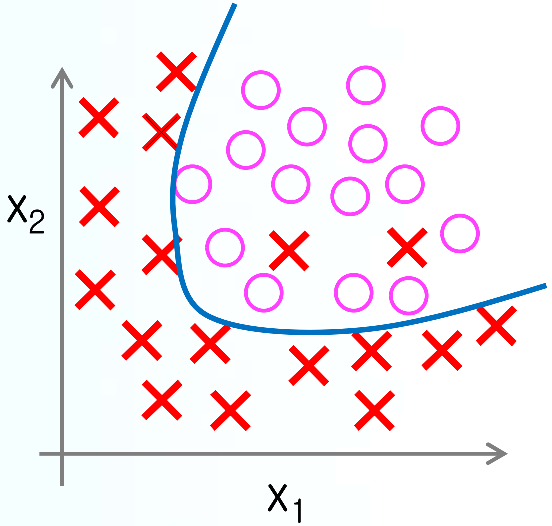
Overfitting

- Example
  - Logistic regression

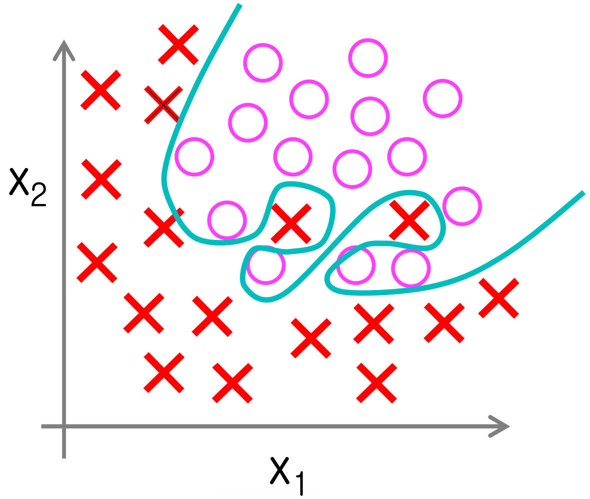


$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

"Underfit"



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

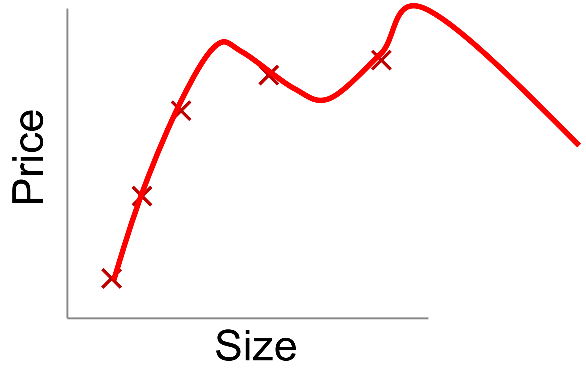


$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$$

"Overfit"

Overfitting

- Addressing overfitting
  - x_1 : size of house
  - x_2 : # of bedrooms
  - x_3 : # of floors
  - x_4 : age of house
  - x_5 : average income in neighborhood
  - x_6 : kitchen size
  - $\vdots$
  - x_{100}



Overfitting

- Addressing overfitting
  - Options:

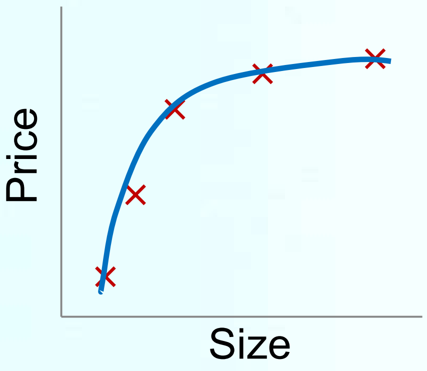
    1) Reduce number of features.
       - Manually select which features to keep.
       - Model selection algorithm

    2) Regularization
       - Keep all the features, but reduce magnitude/values of parameters θ_j .
       - Works well when we have a lot of features,
         each of which contributes a bit to predicting y .

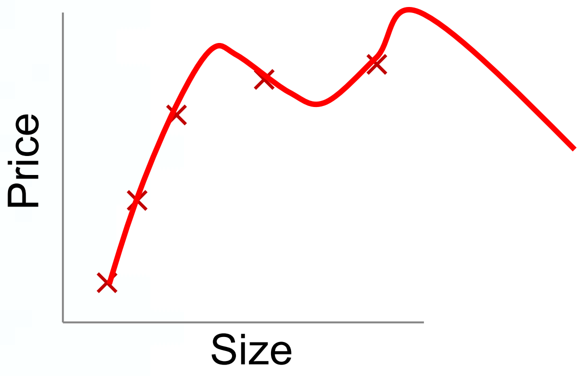
Outline

- Problem of overfitting
- Cost function
- Regularized linear regression
- Regularized logistic regression

Regularization

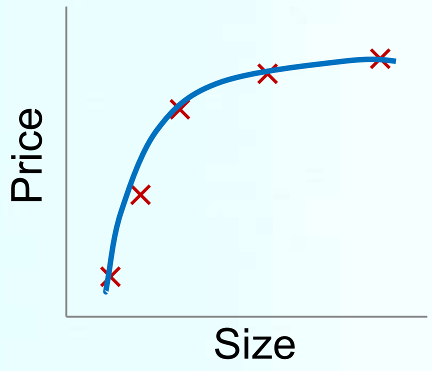


$$\theta_0 + \theta_1 x + \theta_2 x^2$$

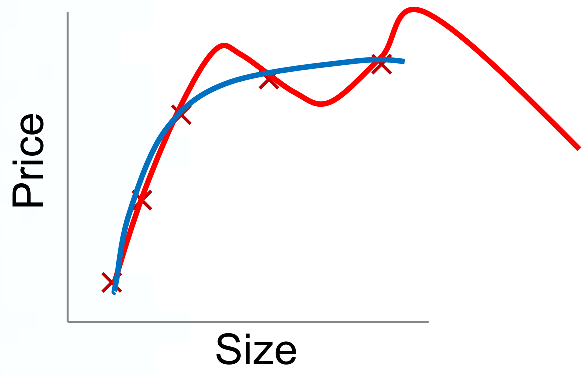


$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Regularization



$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$

- Suppose we penalize and make θ_3, θ_4 really small
  - $\min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right] + 1000\theta_3^2 + 1000\theta_4^2$

➔ $\theta_3 \approx 0,\ \theta_4 \approx 0$

Regularization

- Small values for parameters
  - “Simpler” hypothesis
  - Less prone to overfitting

- Housing
  - Features: $x_1, x_1, \ldots, x_{100}$
  - Parameters: $\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}$

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right] \Rightarrow$$

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

$\boxed{\cancel{\theta_0}}$

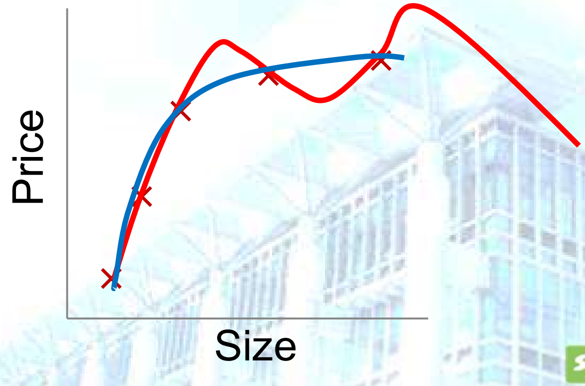
Regularization

- $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$

Regularization parameter $\lambda \geq 0$

- $min_\theta J(\theta)$

- Regularization parameter
  - Controls a trade off btw two goals
    - Fit the training set well
    - Keep parameters small



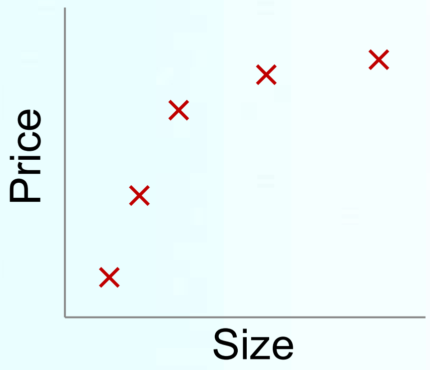
Embedded System Lab.

Regularization

- In regularized linear regression, θ is chosen to minimize

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

- What if λ is set to extremely large value (perhaps for too large for our problem, say $\lambda = 10^{10}$)?

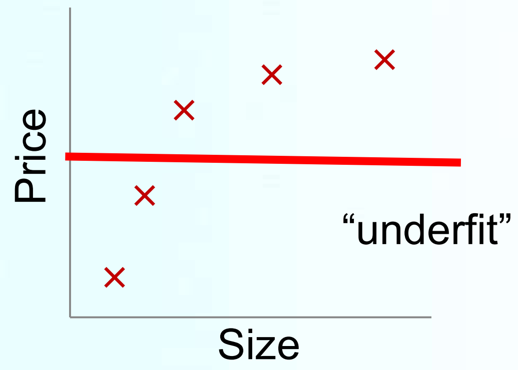


Regularization

- In regularized linear regression, we choose θ to minimize

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$

- What if λ is set to extremely large value (perhaps for too large for our problem, say $\lambda = 10^{10}$)?



$$h_\theta(x) = \theta_0 + \cancel{\theta_1} x + \cancel{\theta_2} x^2 + \cancel{\theta_3} x^3 + \cancel{\theta_4} x^4 \approx \theta_0$$

Outline

- Problem of overfitting
- Cost function
- Regularized linear regression
- Regularized logistic regression

Regularized Linear Regression

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

$$min_\theta J(\theta)$$

Gradient Descent for Linear Regression

- Repeat {

$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)}$$

}

$$(j = 0,1,2, \ldots , n)$$

Gradient Descent for Linear Regression

- Repeat {

$$\theta_0 \leftarrow \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_0^{(i)}$$

$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

} $\quad (j = \cancel{0},1,2,\ldots,n)$

Gradient Descent for Regularized Linear Regression

- Repeat {

$$\frac{\partial}{\partial \theta_0} J(\theta)$$

$$\theta_0 \leftarrow \theta_0 - \alpha \boxed{\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_0^{(i)}}$$

$$\theta_j \leftarrow \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right]$$

} $\qquad (j = \cancel{0}, 1, 2, \ldots, n)$

Gradient Descent for Regularized Linear Regression

- Repeat {

$$\frac{\partial}{\partial \theta_0} J(\theta)$$

$$\theta_0 \leftarrow \theta_0 - \alpha \boxed{\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_0^{(i)}}$$

$$\theta_j \leftarrow \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)}$$

}

$$(j = \cancel{0}, 1, 2, \ldots, n)$$

$$1 - \alpha \frac{\lambda}{m} < 1$$

Gradient Descent for Regularized Linear Regression

- $1 - \alpha\frac{\lambda}{m} < 1$
  - less than 1
  - Usually learning rate α is small and m is large
    - So this typically evaluates to (1 - a small number)
      - $1 - \alpha\frac{\lambda}{m}$ is often around 0.99 to 0.95

- $\theta_j \leftarrow \theta_j(1 - \alpha\frac{\lambda}{m}) - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$
  - If $1 - \alpha\frac{\lambda}{m} = 0.99$
    - θ_j gets multiplied by 0.99 ($\theta_j(1 - \alpha\frac{\lambda}{m})$)
      - Means the squared norm of θ_j a little smaller
  - $-\alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$
    - exactly the same as the original gradient descent

Normal Equation

- If $\theta \in R^{n+1}$
  - $J(\theta_0, \theta_1, \cdots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$
  - $min_\theta J(\theta) = min_\theta \frac{1}{2m}\|X\theta - y\|^2$
    - where $\theta = \begin{bmatrix}\theta_0 \\ \theta_1 \\ \vdots \\ \theta_n\end{bmatrix}$, $X = \begin{bmatrix}\left(x^{(1)}\right)^T \\ \left(x^{(2)}\right)^T \\ \vdots \\ \left(x^{(m)}\right)^T\end{bmatrix}$, $x^{(i)} = \begin{bmatrix}x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)}\end{bmatrix}$, $y = \begin{bmatrix}y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)}\end{bmatrix}$
  - $\theta = (X^TX)^{-1}X^Ty$

Normal Equation for Regularized Linear Regression

- If $\theta \in R^{n+1}$
  - $J(\theta_0, \theta_1, \cdots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

➔ $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$

  - $min_\theta J(\theta) = min_\theta \frac{1}{2m}\left[\|X\theta - y\|^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$

where $\theta = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}$, $X = \underbrace{\begin{bmatrix}\left(x^{(1)}\right)^T\\ \left(x^{(2)}\right)^T\\ \vdots\\ \left(x^{(m)}\right)^T\end{bmatrix}}_{\boldsymbol{m}\times(\boldsymbol{n}+\mathbf{1})}$, $x^{(i)} = \begin{bmatrix}x_0^{(i)}\\ x_1^{(i)}\\ x_2^{(i)}\\ \vdots\\ x_n^{(i)}\end{bmatrix}$, $y = \begin{bmatrix}y^{(1)}\\ y^{(2)}\\ \vdots\\ y^{(m)}\end{bmatrix}$

Normal Equation for Regularized Linear Regression

- If $\theta \in R^{n+1}$
  - $J(\theta_0, \theta_1, \cdots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

    ➔ $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$

  - $min_\theta J(\theta) = min_\theta \frac{1}{2m}\left[\|X\theta - y\|^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$

  - $\frac{\partial}{\partial\theta}J(\theta) = 0$ ➔ $\theta = \left(X^TX + \lambda\underbrace{\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{(\boldsymbol{n}+\mathbf{1})\times(\boldsymbol{n}+\mathbf{1})}\right)^{-1} X^Ty$

Non-Invertibility

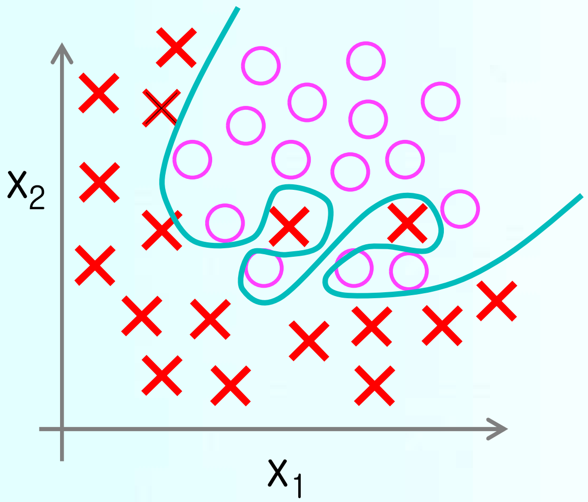
- Suppos $m \le n$
  - (# of examples ≤ # of features)
  - Then, in $\theta = (X^T X)^{-1} X^T y$, $(X^T X)^{-1}$may be singular
  - If λ > 0,

$$\theta = \underbrace{\left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \right)}_{\textbf{Invertible}}{}^{-1} X^T y$$

Outline

- Problem of overfitting
- Cost function
- Regularized linear regression
- Regularized logistic regression

Regularized Logistic Regression

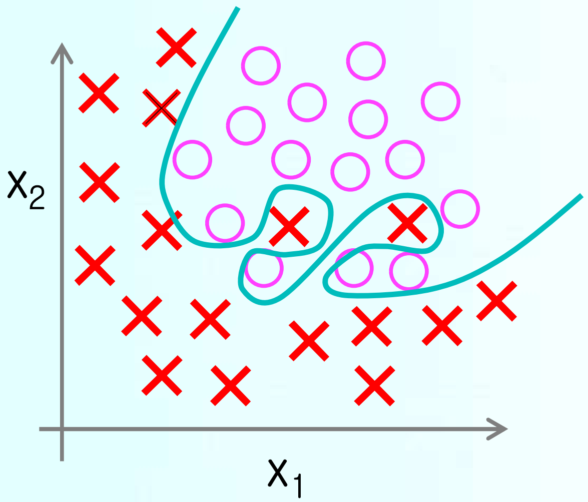


- $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$

Cost function

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log(1 - h_\theta\left(x^{(i)}\right))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Regularized Logistic Regression



The effect of penalizing parameters $\theta_1, \ldots, \theta_{\boldsymbol{n}}$

- $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$

Cost function

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log(1 - h_\theta\left(x^{(i)}\right))\right] + \boxed{\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2}$$

Gradient Descent for Regularized Logistic Regression

- Repeat {

$$\theta_0 \leftarrow \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_0^{(i)}$$

$$\theta_j \leftarrow \theta_j - \alpha \boxed{\left[\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right]}$$

}

$$(j = \cancel{0}, 1, 2, \ldots, n)$$

$$\frac{\partial}{\partial \theta_j} J(\theta)$$

$$h_\theta(z) = \frac{1}{1 + exp(-\theta^T z)}$$

References

- Andrew Ng, https://www.coursera.org/learn/machine-learning
- http://www.holehouse.org/mlclass/07_Regularization.html